ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ МНЕНИЙ И НАСТРОЕНИЙ С ИСПОЛЬЗОВАНИЕМ НЕЧЕТКОЙ ЛОГИЧЕСКОЙ МОДЕЛИ

INTELLIGENT ANALYSIS OF OPINIONS AND MOODS USING FUZZY LOGIC MODEL

A. Tregubov

Summary. Every day, the volume of data suitable for extracting information is growing exponentially. One of the easily accessible sources of such data can be considered social networks. This type of Internet resource has become an excellent platform for the exchange of views, experience and knowledge. This document presents a model based on the theory of fuzzy sets that describes the emotional component of natural language texts. The goal is to test the suitability of the presented model for the automated intellectual analysis of opinions and hidden emotions in the text.

Keywords: fuzzy sets, sentiment analysis, natural language processing, machine learning, subjectivity analysis, knowledge extraction, data analysis, support vector method, maximum entropy method, opinion mining.

Трегубов Артем Сергеевич
Аспирант, Новосибирский национальный исследовательский государственный университет
artem.tregubov@mail.ru

Аннотация. С каждым днем объёмы данных пригодных для извлечения информации растут в геометрической прогрессии. Одним из легкодоступных источников таких данных можно считать социальные сети. Данный тип интернет ресурсов стал прекрасной платформой для обмена мнениями, опытом и знаниями. В данном документе представлена модель, основанная на теории нечетких множеств описывающая эмоциональную составляющую текстов естественного языка. Цель состоит в том, чтобы проверить пригодность представленной модели для автоматизированного интеллектуального анализа мнений и скрытых эмоций в тексте.

Ключевые слова: нечеткие множества, анализ тональности текстов, сентимент-анализ, обработка текстов естественного языка, машинное обучение, анализ субъективности, извлечения знаний, анализ данных, метод опорных векторов, метод максимальной энтропии.

Введение

Сегодня такие сайты как Twitter, Facebook, LinkedIn и т.д. превратились в огромную платформу для обмена знаниями и мнениями в текстовой форме. Эти социальные сети используются для обмена мнениями о различных продуктах, фильмах, политике или о личных предпочтениях. Информация подобного рода очень полезна, из нее можно узнать о современных трендах.

Анализ тональности текста позволяет выявлять скрытые эмоции в тексте. Мнение определяется как кортеж (g, s, h, t), где g — это объект, о котором высказывается мнение, s — мнение, h — субъект, высказывающий мнение, t — время, в которое было высказано данное мнение. Выявление мнений — это процесс, который анализирует и обобщает мнение, выраженное в виде огромных текстовых данных. Анализ настроений классифицирует мнения в разных классах как положительные, отрицательные или нейтральные. Анализ мнений тесно связан с классификацией и выявлением оценок из текстов, представленных на веб-сайтах. Анализ настроений направлен на выявление субъективности и скрытых чувств в тексте. Данная статья предлагает новую модель для выполнения анализа субъективности текста, использующую методы датамайнинга и нечеткой логики.

Хорошо сформулированные мнения могут сыграть важную роль при принятии решений интернет-пользователями. Выявление характеристик — это большая задача, решаемая в рамках сентимент-анализа текста. Каждый объект имеет множество характеристик. Характеристики делятся на явные и неявные. Явные — это характеристики, явно описанные в тексте отзыва. В то время как неявные — это характеристики, затронутые в тексте, но не описанные в нем явно.

Данная статья организована следующим образом: следующая часть дает обзор подходов, описанных в существующей литературе, третья часть посвящена описанию техники классификации и теории нечетких множеств, последняя часть описывает результаты и выводы.

Обзор литературы

Сентимент-анализ — это новая область исследований, в которой были использованы различные техники машинного обучения для выделения и классификации мнений. В рассмотренной литературе были описаны такие методы машинного обучения, как классификатор Байеса [1], метод опорных векторов [2], и нейронные сети [3]. Все эти методы имеют один общий недостаток, присущий такому классу, как методы обучения без учителя — снижение качества результатов на другой доменной модели.

Для эффективного анализа настроений требуется экстраординарный алгоритм выделения признаков. Каждый продукт имеет свой собственный набор функций, а отзывы о нем — это список особенностей продукта, которые являются хорошими индикаторами при классификации отзывов о продуктах.

На этапе извлечения происходит выделение характеристик, а также эмоционально окрашенных слов. Для этого используют алгоритмы обработки текстов на естественном языке.

Подходы на основе лексических правил используют внешние словари, в которых предопределены положительные и отрицательные оценки слов. Это повышает производительность классификатора на основе эмоций. Примерами таких словарей являются WordNet [4], ConceptNet, SenticNet и SentiWordNet. Модели, основанные на аспектах, используются для извлечения аспектов, которые основаны на тематических моделях.

Но текст, созданный пользователями социальных сетей, не структурирован и нелогичен по своей природе. Из литературы было обнаружено, что нечеткие множества [5] эффективны для классификации настроений. Pandey and Goyal [6, 7] использовали нечеткую логику для раннего прогнозирования ошибок программного обеспечения и улучшенной надежности программных систем.

Обоснование исследований

После обзора множества статей было обнаружено достаточное число пробелов в области применения нечеткой логики. Поскольку большинство методов машинного обучения выполняет бинарную классификацию, классифицируя чувства или эмоции как положительные или отрицательные. Также текст, созданный пользователями социальных сетей, имеет сложную структуру. Нечеткие множества способны содержать элементы, которые могут принадлежать разным множествам, а также иметь разную степень принадлежности. Каждое выражение или слово можно классифицировать с помощью нечетких множеств, например, такие слова, как «Отлично» и «Хорошо», получают разную степень принадлежности.

Проблема выделения характеристик является особенно важной в области сентимент-анализа. Для эффективного анализа необходимы экстраординарные подходы к выделению признаков. Каждый продукт имеет свой собственный набор характеристик, и отзывы с четким разделением текста о каждой конкретной функции являются редкостью и представляют большую ценность при составлении датасетов пригодных для обучения [8].

Этап выделения ключевых характеристик подразумевает выделение оцененных характеристик (то есть характеристик продукта, которые были оценены) и слов, содержащих эмоциональную окраску. Обычно для этого используются методы обработки текстов на естественном языке.

В изученной литературе были найдены такие методы машинного обучения с учителем, как классификатор Байеса, Метод максимальной энтропии, метод вспомогательных векторов и нейронные сети, которые наиболее популярны для классификации мнений.

Но тексты, публикуемые в социальных сетях, более неструктурированные [10]. В литературе были найдены методы, использующие нечеткие множества для классификации. Они показали высокие результаты.

После анализа множества статей было выявлено несколько пробелов в развитии сентимент-анализа. Большинство методов подразумевают разделение текстов на 2 класса (положительные и отрицательные). А также в большинстве работ рассматриваются отзывы, размещенные на специализированных сайтах, такие тексты имеют более строгую структуру.

Использование нечетких множеств позволяет обойти первое ограничение и задать разную степень принадлежности отзыва тому или иному множеству.

Основная идея предлагаемого метода подразумевает использование нечетких множеств совместно с методом вспомогательных векторов.

Исследование

Методы глубокого обучения подразумевают наличие большого объема информации для проведения обучения, а также знание основ выявления паттернов и знаний. В литературе описаны несколько подходов, таких как регрессия, классификация, ассоциация и кластеризация, которые используются для классификации данных и выявления значимых характеристик.

Также в литературе описаны техники классификации: K-средних, ANN, метод вспомогательных векторов. Последние показывают наиболее высокие результаты в области обработки текстов при достаточно высокой производительности [8].

Процесс интеллектуального анализа мнений подразумевает определение, выделение и классификацию эмоциональной составляющей, скрытых в текстовых сообщениях. Процесс интеллектуального анализа состоит из следующих этапов:

1. Сбор данных.
2. Обработка данных.
3. Обобщение обработанных результатов.

Обработка данных в свою очередь состоит из следующих этапов:

1. Извлечение признаков.
2. Классификация чувств.

Обобщение результатов выполняется с целью получить результаты, сгруппированные по различным признакам, которые будут понятны и легко воспринимаемы человеком. Дальнейший анализ подразумевает работу пользователя со сгруппированными результатами.

Теория нечетких множеств

В классической теории множеств принадлежность элементов множеству оценивается в бинарных терминах в соответствии с чётким условием. Считается, что элемент x принадлежит множеству X, если выполнено некоторое условие истинно. Данное условие не может быть выполнено в некоторой степени.

Основная идея теории нечетких множеств заключается в том, что истинность может быть выражена не только лишь 2 значениями, а напротив, любым значением между 0 и 1, тем самым показывая степень истинности.

Пусть $\widetilde{\mathrm{A}}$ — это нечеткое множество множества A, тогда степень принадлежности элемента χ из U определяется функцией $\mu\widetilde{\mathrm{A}}(\chi)$, называемой функцией принадлежности:

$$\widetilde{\mathrm{A}} = \left\{\frac{\mu_1}{\chi_1} + \frac{\mu_2}{\chi_2} + \frac{\mu_3}{\chi_n} + \cdots + \frac{\mu_n}{\chi_n}\right\}$$

где μ — степени принадлежности элементов нечеткому множеству, χ — элементы множества U.

Разработка модуля обработки текста с использованием нечеткой логики

Полученные тексты отзывов должны быть предварительно обработаны системой для выявления ключевых слов, отражающих субъективное отношение к функциям. Затем полученные результаты отдаются в модуль для обработки с использованием нечеткой логики, которые бы возвращали степень принадлежности к некоторым множествам.

Входными значениями являются мнения о свойствах, разделенные на 5 основных категорий (от очень плохого до очень хорошего). Диапазоны значения для них определяются по формуле:

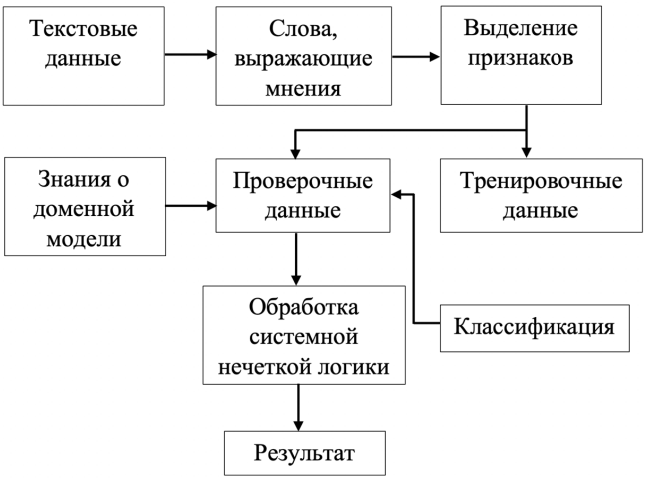

Рис. 1

$$\left[1 - \frac{\{\log_{10}(1{:}5)\}}{\log_{10}(5)}\right]$$

Результатом расчетов является степень удовлетворенности каждой конкретной характеристики.

Предлагаемое решение

Архитектура представлена на картинке ниже. Для обработки используются техники анализа данных, с использованием языка R и Matlab. На каждом этапе, результаты работы текущего модуля передаются в качестве входных данных для следующего модуля (рис. 1).

Данные для обработки

В качестве источника данных была использована социальная сеть Twitter. С помощью языка R были выделены записи, относящиеся к конкретному продукту и его отдельным свойствам.

Предварительная обработка данных

В этом разделе обсуждаются этапы предварительной обработки, используемые в предлагаемой модели. В начале просматриваются все тексты, совпадающие с ключевыми словами, фиксируются в корпусе, разбиваются на токены и стоп-слова. Для извлечения выражений, использующихся для выражения мнения, применяются методы обработки текстов на естественном языке. Так же строится матрица терминов документов.

Извлечение свойств

Особенности объектов извлекаются в каждом предложении, а субъективные признаки классифицируются.

Тацблица 1

Алгоритм	Точность	Полнота	F-мера
SVMM	0.22	1	0.36
MAX	0.20	1	0.33

Мнения, содержащие слова, извлекаются и классифицируются как положительные и отрицательные, а их степень полярности определяется с помощью нечетких множеств.

Обработка данных

В предложенной модели для классификации настроений используются алгоритмы обучения с учителем, вспомогательные векторные машины (SVM) и метод максимальной энтропии.

Классификация мнений

Модули языка R используются для классификации. Модель обучения строится с использованием следующих шагов:

1. Создать матрицу терминов для документа.
2. Создать контейнер.
3. Создать модель, загрузив контейнер в алгоритм машинного обучения.
4. Протестировать модель.

Результаты

Результаты были основаны на наборе данных, который уже упоминался в разделе 4, и мы проанализировали различные методы классификации. Было проведено сравнение результатов полученных на одинаковом наборе данных и результаты описаны в таблице 1.

Также были сравнены параметры, такие как точность классификации и время обучения.

Выявлено, что SVM обладает наиболее высокой точностью и значительно более высокой скоростью обучения.

Предложенные методы могут быть использованы для анализа отзывов пользователей и отзывов, размещенных в социальных сетях, для анализа товаров и услуг, с целью улучшения качества и как следствие повышения лояльности покупателей.

ЛИТЕРАТУРА

1. Dinu, Liviu P., and Iulia Iuga. 2012. The Naive Bayes Classifier in Opinion Mining. In Search of the Best Feature Set. Berlin: Springer.
2. T. Wang, H. Huang, S. Tian, J. Xu, Feature Selection for SVM via Optimization of Kernel Polarization with Gaussian ARD Kernels. Expert Systems with Applications, 37(9) (2010) 6663–6668
3. L.S. Chen, H. J. Chiu, Developing a Neural Network based Index for Sentiment Classification. Proceedings of the International MultiConference of Engineers and Computer Scientists (2009) 744–749.
4. George A. Miller. 1995. WordNet: A Lexical Database for English.Commun. ACM38, 11 (1995), 39–41.
5. Zadeh, L.A. 1965. Fuzzy Sets. Information and Control 8: 338–353. 20.
6. Pandey, A.K., and N. K. Goyal. 2009. A Fuzzy Model for Early Software Fault Prediction Using Process Maturity and Software Metrics. International Journal of Electronics Engineering 239–245. 21.
7. Pandey, A.K., and N. K. Goyal. 2010. Predicting Fault-prone Software Module Using Data Mining Technique and Fuzzy Logic. International Conference.
8. M. Abdel Fattah, New term weighting schemes with combination of multiple classifiers. Neurocomputing,167 (2015) 434–442.
9. L. Gui, Y. Zhou, R. Xu, Y. He, Q. Lu, Learning representations from heterogeneous network for sentiment classification of product reviews. Knowledge-Based Systems, 124 (2017) 34–45.
10. Emitza Guzman, Rana Alkadhi, and Norbert Sey. 2017. An exploratory study ofTwitter messages about software applications.Requirements Engineering22, 3(2017), 387–412.

Журнал «Современная наука: актуальные проблемы теории и практики»